Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Module 4 – Advanced Analytics - Theory and Methods

Module 4: Advanced Analytics – Theory and Methods

Part 4: Logistic Regression

During this Part the following topics are covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to Choose (+) and Cautions (-) of the logistic regression model

Logistic Regression

- Used to estimate the probability that an event will occur as a function of other variables
  - The probability that a borrower will default as a function of his credit score, income, the size of the loan, and his existing debts
- Can be considered a classifier, as well
  - Assign the class label with the highest probability
- Input variables can be continuous or discrete
- Output:

A linear expression for predicting the log-odds ratio of outcome as a function of drivers. (Binary classification case)

- Log-odds ratio easily converted to the probability of the outcome

$$odds\_ratio = \frac{P(x=true)}{P(x=false)} = \frac{\mathrm{P(x=true)}}{1-\mathrm{P(x=true)}}$$

Logistic Regression Use Cases

- The preferred method for many binary classification problems:
  - Especially if you are interested in **the probability of an event, not just predicting "yes or no“**
  - Try this first; if it fails, then try something more complicated
- Binary Classification examples:
  - The probability that a borrower will default
  - The probability that a customer will churn
- Multi-class example
  - The probability that a politician will vote yes/vote no/not show up to vote on a given bill

Logistic Regression Model - Example

$$\text{default} = f(\text{creditScore}, \text{income}, \text{loanAmt}, \text{existingDebt})$$

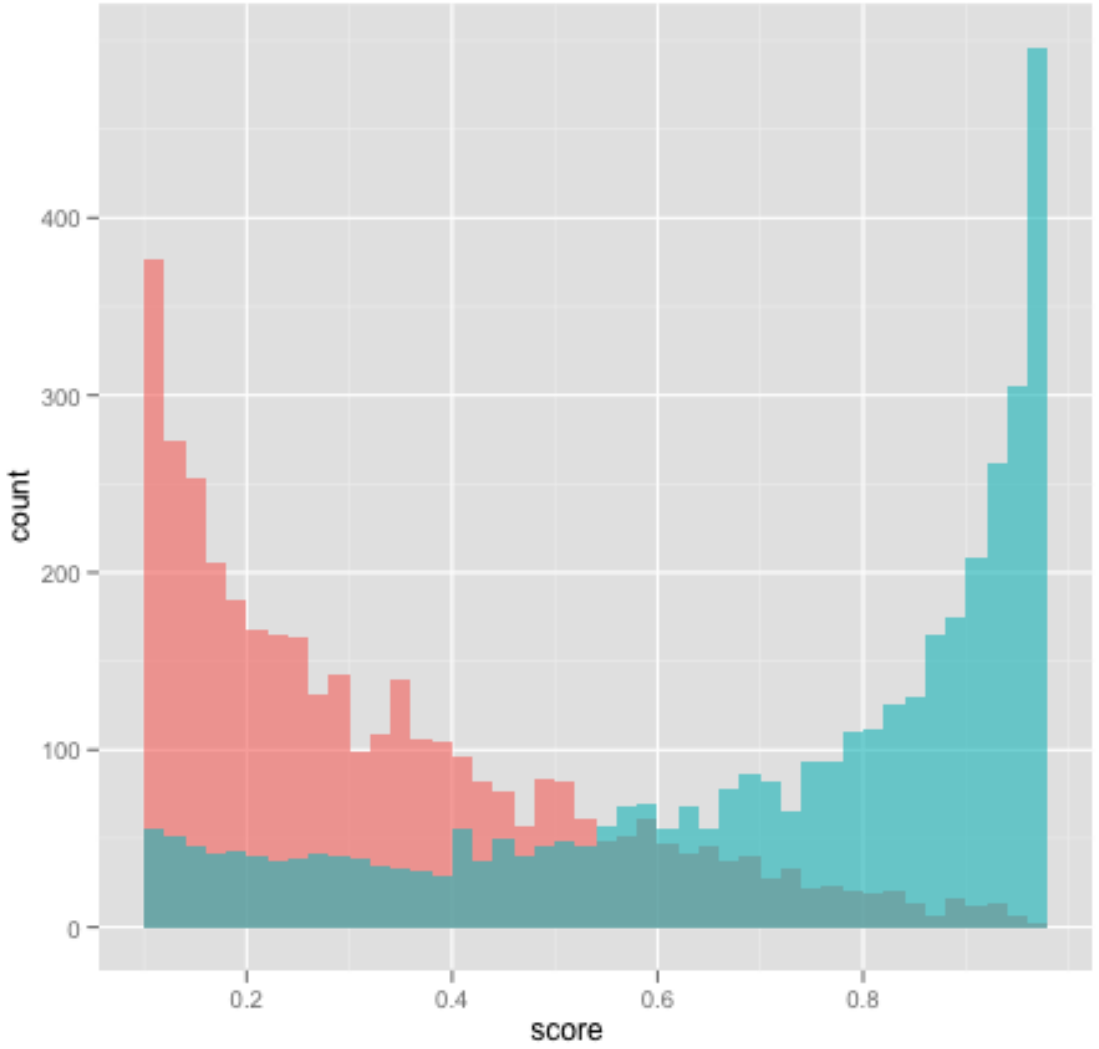
- Training data: default is 0/1
  - default=1 if loan defaulted
- The model will return the probability that a loan with given characteristics (value for each of the input variables) will default
- If you only want a "yes/no" answer, you need a threshold
  - The standard threshold is 0.5

Logistic Regression- Visualizing the Model

Overall fraction of default: ~20%

Logistic regression returns a score that estimates the probability that a borrower will default

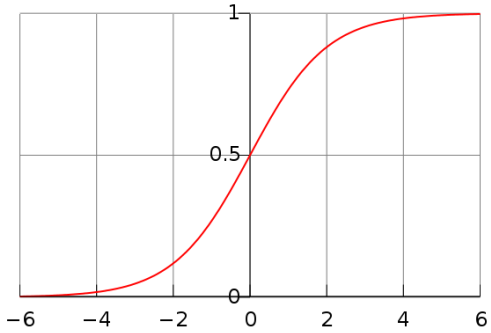
The graph compares the distribution of defaulters and non-defaulters as a function of the model's predicted probability, for borrowers scoring higher than 0.1



Blue=defaulters

Technical Description (Binary Case)

$$ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1}$$



- y=1 is the case of interest: 'TRUE'
- LHS is called logit(P(y=1))
  - hence, "logistic regression"
- logit(P(y=1)) is inverted by the sigmoid function
  - standard packages can return probability for you
- Categorical variables are expanded as with linear regression
- Iterative solution to obtain coefficient estimates, denoted b_j
  - "Iteratively re-weighted least squares"

Interpreting the Estimated Coefficients, b_i

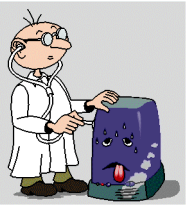
- Invert the logit expression:

$$\frac{P(y=1)}{1-P(y=1)} = \exp\left(\sum_{j=0}^{K} b_j x_j\right)$$

$$= \prod_{j=0}^{K} \exp(b_j x_j)$$

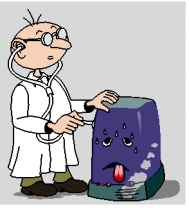
- $\exp(b_j)$ tells us how the odds-ratio of y=1 changes for every unit change in $x_{j\ldots\ldots}$ *It gives annotation of how each variable affects the outcome*
- Example: $b_{creditScore} = -0.69$
  - $\exp(b_{creditScore}) = 0.5 = 1/2$
  - for the same income, loan, and existing debt, the odds-ratio of default is halved for every point increase in credit score
- Standard packages return the significance of the coefficients in the same way as in linear regression

Diagnostics



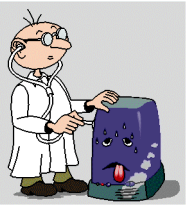
- Hold-out data:
  - Does the model predict well on data it hasn't seen?
- N-fold cross-validation: Formal estimate of generalization error
- "Pseudo-R^2" : 1 – (deviance/null deviance)
- Pseudo R^2 which is **a measure of how well the model explains the data.**
  - Deviance, null deviance both reported by most standard packages
  - The fraction of "variance" that is explained by the model
  - Used the way R^2 is used

Diagnostics (Cont.)



- Check the coefficients
  - Do the signs make sense? Are the coefficients excessively large?
    - Wrong sign is an indication of correlated inputs, but doesn't necessarily affect predictive power.
    - Excessively large coefficient magnitudes may indicate strongly correlated inputs; you may want to consider eliminating some variables, or using regularized regression techniques.
    - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output (and doesn't predict well on the rest).
      - Try a Decision Tree on that variable, to see if you should segment the data before regressing.

Diagnostics: ROC Curve

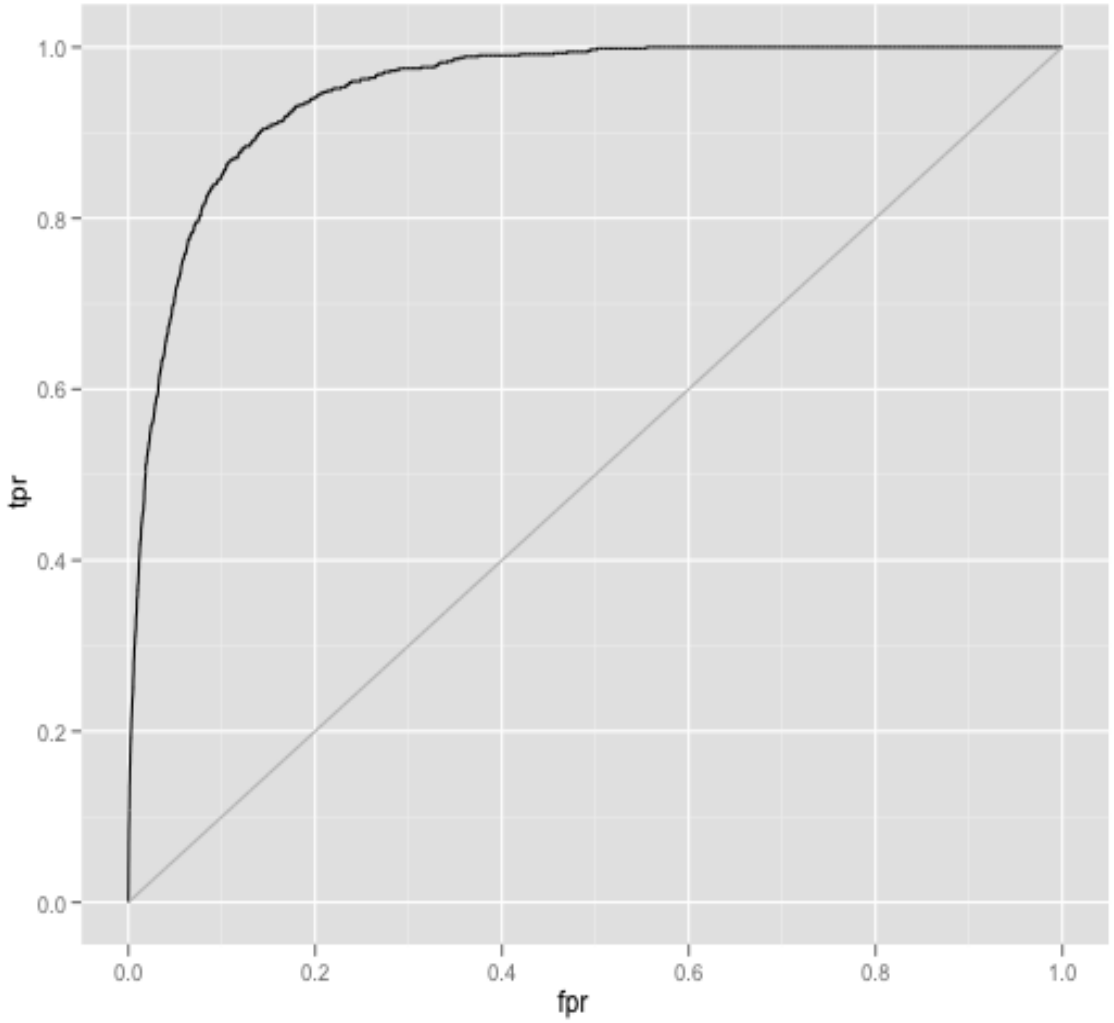


$$\text{FPR} = \frac{\#\text{ false positives}}{\text{all negatives}}$$

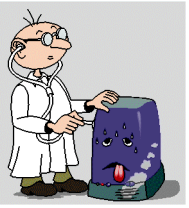
$$\text{TPR} = \frac{\#\text{ true positives}}{\text{all positives}}$$

Area under the curve (AUC) tells you how well the model predicts. (Ideal AUC = 1)

For logistic regression, ROC curve can help set classifier threshold

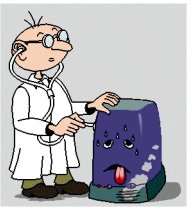


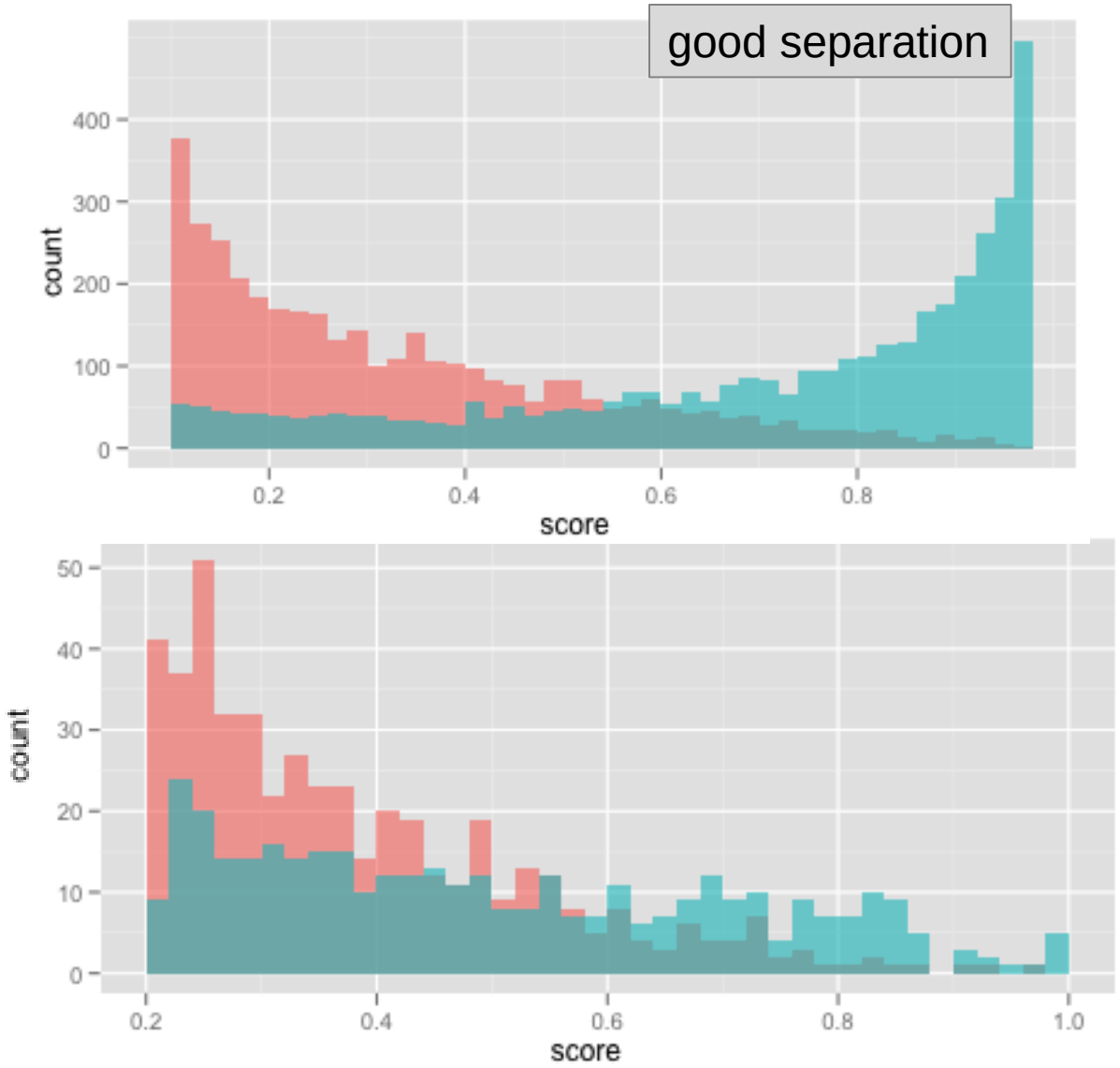
Diagnostics: ROC Curve



- A receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.
- The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The true-positive rate is also known as sensitivity, or recall in machine learning.
- The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity).
- The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification.

Diagnostics: Plot the Histograms of Scores





Logistic Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Explanatory value: Relative impact of each variable on the outcome in a more complicated way than linear regression	Does not handle missing values well
Robust with redundant variables, correlated variables With correlated variables, the prediction is not impacted but we lose some explanatory value with the fitted model.	Assumes that each variable affects the log-odds of the outcome linearly and additively Variable transformations and modeling variable interactions can address this to some extent
Concise representation of the outcome with the coefficients	Cannot handle variables that affect the outcome in a discontinuous way. Step functions
Easy to score data with the coefficients	Doesn't work well with discrete drivers that have a lot of distinct values For example, ZIP code
Returns good probability estimates of an event	
Preserves the summary statistics of the training data "The probabilities equal the counts"	

Check Your Knowledge

Your Thoughts?

1. What is a logit and how do we compute class probabilities from the logit?
2. How is ROC curve used to diagnose the effectiveness of the logistic regression model?
3. What is Pseudo R^2 and what does it measure in a logistic regression model?
4. How do you describe a binary class problem?
5. Compare and contrast linear and logistic regression methods.

Module 4: Advanced Analytics – Theory and Methods

Part 4: Logistic Regression - Summary

During this Part the following topics were covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to Choose (+) and Cautions (-) of the logistic regression model

Lab Exercise 7: Logistic Regression

This Lab is designed to investigate and practice Logistic Regression.

After completing the tasks in this lab you should be able to:

- Use R functions for Logistic Regression – (*also known as Logit*)
- Predict the dependent variables based on the model
- Investigate different statistical parameter tests that measure the effectiveness of the model

Lab Exercise 7: Logistic Regression - Workflow

1. Define the problem and review input data
2. Set the Working Directory
3. Read in and examine the data
4. Build and review logistic regression model
5. Review the results and interpret the coefficients
6. Visualize the model using the Plot function
7. Use Relevel function to re-level the Price factor with value 30 as the base reference
8. Plot the ROC curve
9. Predict Outcome given Age and Income
10. Predict outcome for a sequence of Age values at price 30 and mean income
11. Predict outcome for a sequence of income at price 30 and mean age
12. Use logistic regression as a classifier

Thanks